Statistica Sinica **11**(2001), 827-842

ON POSTERIOR CONSISTENCY IN SELECTION MODELS

Jaeyong Lee

Pennsylvania State University

Abstract: Selection models are appropriate when the probability that a potential datum enters the sample is a nondecreasing function of the numeric value of the datum. It is rarely justifiable to model this function, called the weight function, with a specific parametric form, but appealing to model it with a nonparametric prior centered around a parametric form. The Bayesian analysis with a Dirichlet process prior for the weight function is considered and it is proved that the posterior is consistent under the weak topology.

Key words and phrases: Dirichlet process, posterior consistency, selection model, weight function.

1. Introduction

Suppose $X_1, \ldots, X_N$ are independent and identically distributed (i.i.d.) with density $f(x)$. In many situations a potential datum, X_i, enters the sample only with some probability $w(X_i)$, typically a nondecreasing function of the numerical value of the datum; this function is called the weight function and will be assumed to be nondecreasing throughout the paper.

Selection models arise naturally in many practical applications, such as oil discovery (Meisner and Demirmen (1981), Nair and Wang (1989), West (1994, 1996)), aerial survey (Cook and Martin (1974), Patil and Rao (1977), Sun and Woodroofe (1997), and Lee and Berger (1999)), meta-analysis (Iyengar and Greenhouse (1988), Bayarri and DeGroot (1992), and Silliman (1997a, b)), econometrics (Maddala (1977)), and astronomy (Sun and Woodroofe (1997)).

A nondecreasing weight function selects observations that overrepresent large values and it is crucial to take the selection mechanism into consideration. In most cases, the exact form of weight function is not known and a specific parametric form can rarely be justified. Hence a Bayesian approach with a nonparametric prior centered around a parametric form of the weight function is appealing. Lee and Berger (1999) developed such an approach, showing how it can be computationally implemented via Markov chain Monte Carlo (MCMC). They also observed an interesting phenomenon, termed *practical nonidentifiability*, wherein the likelihood has two modes, one around the true parameter value and the other representing a near constant weight function. While this complicates inference by frequentists and Bayesians, a solution has been proposed by Lee and Berger (1999). Theoretically, however, this phenomenon increases concern as to whether the posterior is consistent in selection models. As shown by Diaconis and Freedman (1986), in a nonparametric location problem with a Dirichlet process prior on the error distribution, the posterior can be inconsistent. Subsequent papers studying consistency include Diaconis and Freedman (1993), Cox (1993), Barron, Schervish and Wasserman (1999), and Ghosal, Ghosh and Ramamoorthi (1999).

In this paper, we establish posterior consistency for selection models of the type studied in Lee and Berger (1999). An important technical complication arises because the observations in selection models are not i.i.d., while most of the theorems (see Schwartz (1965), Barron (1988), Barron, Schervish and Wasserman (1999), and Ghosal, Ghosh and Ramamoorthi (1999)) assume they are. We resolve this issue by using a result from Barron (1988), with the observation that the posterior arising from a selection model can be represented as one arising from i.i.d. observations with varying priors. A second unusual aspect of the problem is that the number of unobserved observations is unknown. In studying consistency, we need to let this number go to infinity, so the prior distribution must "allow" this to happen in a suitable sense.

2. Notation and Models

2.1. Introduction

Let $X = \{X_1, \ldots, X_N\}$ be an i.i.d. sample from the density $f(x|\theta)$ with respect to Lebesgue measure (or counting measure), with support $\mathcal{X} \subset R$ and θ unknown in Θ, where Θ is a subset of an Euclidean space, and let $0 \le w(x) \le 1$ be a nondecreasing weight function. Each X_i is observed (or selected) with probability $w(X_i)$. The overall probability of selection is $v(\theta, w) = \int w(x)f(x|\theta)dx$. Let O be the observed portion of X, written $\{x_{i_1}, \ldots, x_{i_n}\}$, and $M = X \backslash O$ be the portion of X that is not selected. Thus n is the number of selected observations. The conditional density of X_i, given that it is selected, is

$$f^w(x|\theta) = \frac{w(x)f(x|\theta)}{v(\theta, w)}. \tag{1}$$

Any bounded non-trivial $w(x)$ can be renormalized so as to have supremum equal to one, without affecting (1). We assume this has been done, so $w(x)$ can be viewed as an element of the space, $\mathcal{W}$, of cumulative distribution functions (cdfs) on $\mathcal{X}$.

If both the weight function w and density f are modeled nonparametrically, the model becomes unidentifiable. In this paper, w is modeled nonparametrically and f is given parametric form. For more detailed discussion of identifiability, see Lee and Berger (1999).

2.2. Sampling plans

We consider two common mechanisms by which the data are generated, the binomial and negative binomial sampling plans. Under the binomial sampling plan, N is fixed beforehand (though possibly unknown), $X_1, \ldots, X_N$ are drawn from a density $f(x|\theta)$, and each X_i enters the sample with probability $w(X_i)$. The density of $O = \{x_{i_1}, \ldots, x_{i_n}\}$ and n is then

$$p(O, n|N, w, \theta) = \binom{N}{n} v(\theta, w)^n (1 - v(\theta, w))^{N-n} \prod_{j=1}^{n} \frac{f(x_{i_j}|\theta) w(x_{i_j})}{v(\theta, w)}.$$

Under the negative binomial sampling plan, n is fixed beforehand, $X_1, X_2, \ldots$ are sampled from the population until n of them are selected in the sample, and the density of O and N is

$$p(O, N|n, w, \theta) = \binom{N-1}{n-1} v(\theta, w)^n (1 - v(\theta, w))^{N-n} \prod_{j=1}^{n} \frac{f(x_{i_j}|\theta) w(x_{i_j})}{v(\theta, w)}.$$

Note that, under both plans, n is known to the statistician after sampling, but N could be either known or unknown. If N is known, the sampling plans are equivalent for Bayesian analysis because their likelihoods are proportional. If N is unknown, however, the likelihoods are different. Under the negative binomial N is a part of the sampling plan and hence a prior on N is not necessary, while under the binomial with unknown N a prior on N is necessary. Interestingly, in the latter situation, the improper prior $\pi(N) \propto N^{-1}$ results in the same posterior as obtained under negative binomial sampling (Bayarri and DeGroot (1990)).

2.3. The prior distribution

Prior distributions need to be specified for θ, the unknown parameter in the density $f(\cdot|\theta)$, for the unknown w, and for the unknown N in the binomial sampling plan. A proper prior is considered for θ.

The prior we consider for the weight function, w, is based on the Dirichlet process (Ferguson (1973)) D_α, with parameter α, where α is a nonnull finite measure on $\mathcal{X}$. This is a prior on the space of probability measures on $\mathcal{X}$ such that, for every Borel measurable partition $(B_1, \ldots, B_k)$, the random element P from the process satisfies

$$(P(B_1), \ldots, P(B_k)) \sim D(\alpha(B_1), \ldots, \alpha(B_k)),$$

where $D(\alpha_1, \ldots, \alpha_k)$ denotes the Dirichlet distribution with parameters, α_1, ..., α_k. The parameter (or base measure) of the Dirichlet process is chosen to be of the form $\alpha(\cdot|\eta)$. The mean of this Dirichlet process is $\alpha(\cdot|\eta)/\alpha(\mathcal{X}|\eta)$, which we thus choose to be the subjective guess as to the parametric form of the weight function, with η representing a possibly unknown parameter from this parametric form. The parameter η, of the centering parametric form, will be assigned a proper prior density $\mu(\eta)$.

As discussed in Section 2.2, a prior on N is necessary if N is unknown under the binomial sampling plan. In the unknown N case, much of the uncertainty of the posterior comes from that of N; hence in practice, use of a proper subjective prior for N is to be encouraged. For investigating consistency, however, many proper priors on N are unsuitable, since consistency must be studied as $N \to \infty$; clearly, for instance, a prior on N with bounded support could not be used. Similarly priors on N that have too sharp tails cannot sensibly accommodate $N \to \infty$. We thus consider priors with a polynomial tail, i.e.,

$$\pi_1(N) \propto \begin{cases} \frac{1}{N^\gamma}, & \text{if } N \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

where, $\gamma \geq 1$. Finally, θ, w and N are taken to be independent.

3. Main Results

Suppose the parametric family of cdfs $F(\cdot|\theta)$, $\theta \in \Theta$ (an open subset of an Euclidean space), has the same support $\mathcal{X}$ (either (a, ∞) with $-\infty \leq a < \infty$, the continuous case, or $\{a, a+1, \ldots\}$ with a an integer, the discrete case). Denote by $f(\cdot|\theta)$ the density of $F(\cdot|\theta)$ with respect to a measure ν (either Lebesgue measure, the continuous case, or the counting measure, the discrete case). Let $\mathcal{W}$ be the space of all distribution functions on $\mathcal{X}$ and define $\mathcal{P} = \{f^w(\cdot|\theta) : \theta \in \Theta, w \in \mathcal{W}\}$. Let w_0 and θ_0 be the true weight function and parameter value, and let $\inf \mathcal{X} = a$.

Definition 1. The posterior $\pi(\cdot|O, n)$ is said to be consistent at $P_0 = f^{w_0}(\cdot|\theta_0) \in \mathcal{P}$ if $\pi(U|O, n) \to 1$, $P_0 - a.s$, for any weak neighborhood U of P_0. The limit is taken as $n \to \infty$ under the negative binomial sampling plan, and as $N_0 \to \infty$ under the binomial sampling plan.

Remark. The weak neighborhoods are open neighborhoods defined by the weak topology on the space of probability measures induced by weak convergence.

General Assumptions

G1 : The base measure of the Dirichlet process, α, has support $\mathcal{X}$. With a slight notational abuse, $\alpha(t)$ denotes $\alpha(\inf \mathcal{X}, t]$.

G2 : The prior $\pi(\theta)$ on Θ assigns a positive mass to every nonempty open subset of Θ.

G3 : The two maps from Θ to R,

$$\theta \rightarrow \int_{\mathcal{X}} f(x|\theta_0)|\log \frac{f(x|\theta_0)}{f(x|\theta)}|d\nu(x)$$
$$\theta \rightarrow \int_{\mathcal{X}} |f(x|\theta_0) - f(x|\theta)|d\nu(x)$$

are continuous at $\theta = \theta_0$.

Additional Assumptions for $\mathcal{X} = (a, \infty)$
C1 : $w_0(\cdot)$ is continuous on $\mathcal{X}$.
C2 : $\int_{\mathcal{X}} f(x|\theta_0)|\log w_0(x)|dx < \infty$.
C3 : $\lim_{t\rightarrow \inf \mathcal{X}} F(t|\theta_0)|\log w_0(t)| = 0$.
C4 : $\int_{\mathcal{X}} f(x|\theta_0)/\alpha^2(x)dx < \infty$.
Note that C2 and the assumption that $\mathcal{X}$ is the support of $f(\cdot|\theta)$ for all θ imply that $w_0(x) > 0$ for all $x \in \mathcal{X}$.

Additional Assumption for $\mathcal{X} = \{a, a+1, \ldots\}$
D1 : $w_0(a) > 0$.

Theorem 1. *If* $\mathcal{X} = (a, \infty)$*, the posterior is consistent at* P_0 *under assumptions* $G1 - G3$ *and* $C1 - C4$.

Theorem 2. *If* $\mathcal{X} = \{a, a+1, \ldots\}$*, the posterior is consistent at* P_0 *under assumptions* $G1 - G3$ *and* $D1$.

Theorems 1 and 2 are for the case in which the Dirichlet process prior has a specified base measure α. The following theorem deals with the case of a mixture Dirichlet process, centered on a parametric family.

Theorem 3. *Let* $w \sim D_{\alpha(\cdot|\eta)}$ *and* $\eta \sim \mu(\cdot)$*, where* μ *is a proper prior on the hyperparameter* η*. Then the conclusions of Theorem* 1 *and* 2 *remain valid if their assumptions are true with positive* μ *probability.*

Normal Example. The underlying density is $N(\mu, \sigma^2)$ and the priors on the parameters are as follows: $w \sim D_{A\cdot N(a,b^2)}$, $\mu|\sigma^2 \sim N(\mu^m, \sigma^2/k^m)$, $\sigma^{-2} \sim Gamma(\alpha^m, \beta^m)$, $N \sim \pi_1$, where $\mu^m \in R$ and A, k^m, α^m, $\beta^m > 0$.

Suppose C1-C3 hold for $\theta_0 = (\mu_0, \sigma_0^2)$ and w_0. It is straightforward to verify G1-G3. Assumption C4 holds only when $b^2 > 2\sigma_0^2$, hence, Theorem 1 does not guarantee the consistency for all normal base measures of Dirichlet process priors. However with a hyperprior $a|b^2 \sim N(\mu^w, b^2/k^w)$, $b^{-2} \sim Gamma(\alpha^w, \beta^w)$, where $\mu^w \in R$ and $k^w, \alpha^w, \beta^w > 0$, the posterior is consistent by Theorem 3, because the set $\{b^2 \in R : b^2 > 2\sigma_0^2\}$ has positive prior probability.

Poisson Example. Suppose X has a Poisson distribution with mean λ and the priors on the parameters are as follows: $w \sim D_{A\cdot Geometric(p)}$, $\lambda \sim Gamma(\alpha, \beta)$,

$N \sim \pi_1$, where A, α, β are positive constants. Assumptions G1-G3 are easy to verify. Under assumption D1, the posterior is consistent. By Theorem 3, with a hyperprior $p \sim Beta(a, b)$ with $a, b > 0$, the posterior is still consistent.

It should be noted that, even though the posterior is consistent on $\mathcal{P}$, the theorems do not say that the posterior is consistent on the parameter space $\Theta \times \mathcal{W}$. If the inverse of the map T from $\Theta \times \mathcal{W}$ to $\mathcal{P}$, $T : (\theta, w) \longrightarrow f^w(\cdot|\theta)$, is continuous at $f^{w_0}(\cdot|\theta_0)$, the posterior would be consistent on the parameter space $\Theta \times \mathcal{W}$ by the *continous mapping theorem*. This is not usually the case.

4. Proofs of Results

We utilize the result of Schwartz (1965), as discussed by Barron (1986), for the case of a negative binomial sampling scheme. For the binomial sampling scheme, Schwartz's theorem cannot be applied because observations are not i.i.d., so we utilize a result due to Barron (1988). The exact form of Barron's theorem cited here does not appear in his paper, but is a consequence of his Theorem 5 and Lemma 8. Schwartz's theorem is a corollary of Barron's theorem.

Theorem 4. (*Barron*) *Let $U_1, U_2, \ldots$ be i.i.d. random variables with common distribution Q. Suppose Q belongs to $\mathcal{P}$, a family of probability measures dominated by a σ-finite measure ν. Suppose a sequence of priors $\{\pi_n\}$ puts positive mass on every Kullback-Leibler ball K_δ around Q which is not exponentially small, i.e., for every $\delta, r > 0$, there exists an n_0 such that for all $n > n_0$, $\pi_n\{P : I(Q, P) < \delta\} > e^{-nr}$, where $I(Q, P) = \int q \log(q/p) d\nu$ and q and p are the densities of Q and P with respect to ν. Then the posterior is consistent at Q.*

Theorem 5. (*Schwartz*) *The posterior is consistent at Q under the assumptions of Barron's theorem with the sequence of priors $\{\pi_n\}$ replaced by a fixed prior π which puts positive mass on every Kullback-Leibler ball around Q.*

We begin with a series of technical Lemmas. The first lemma is from Sethuraman (1983). For a finite measure μ on $\mathcal{X}$ and a measurable set A in $\mathcal{X}$, let $\mu|_A(B) = \mu(A \cap B)/\mu(A)$, for all measurable sets B in $\mathcal{X}$.

Lemma 1. (*Projection Lemma*) *Suppose $P \sim D_\alpha$ and A is a measurable set. Then $P|_A$, $P|_{A^c}$ and $P(A)$ are independent and their distributions are $D_{\alpha(A)\cdot\alpha|_A}$, $D_{\alpha(A^c)\cdot\alpha|_{A^c}}$ and $Beta(\alpha(A), \alpha(A^c))$, respectively.*

Lemma 2. *Suppose $\mathcal{X} = \{a, a+1, \ldots\}$ and G1 holds. Then, for any $\epsilon > 0$, $D_\alpha\{w : \sup_{x\in\mathcal{X}} |w_0(x) - w(x)| < \epsilon\} > 0$.*

Proof. Let $\epsilon > 0$ be given. Take $b \in \mathcal{X}$ so that $w_0(b) > 1-\epsilon/4$. Since $(w(a), w(a+1) - w(a), \ldots, w(b) - w(b-1), 1 - w(b))$ follows a finite-dimensional Dirichlet distribution with parameters $(\alpha(a)$, $\alpha(a+1) - \alpha(a)$, $\ldots$, $\alpha(b) - \alpha(b-1)$, $\alpha(\mathcal{X}) -$

$\alpha(b))$, $D_\alpha(A) > 0$, where $A = \{w : |w(x) - w_0(x)| < \epsilon/4, x = a, a+1, \ldots, b\}$. If $w \in A$ then $w(b) > 1 - \epsilon/2$, since $w_0(b) > 1 - \epsilon/4$ and $|w(b) - w_0(b)| < \epsilon/4$. This implies, for $x = b, b+1, \ldots,$ that

$$|w(x) - w_0(x)| \le |1 - w(x)| + |1 - w_0(x)| < \epsilon/2 + \epsilon/2 = \epsilon.$$

Hence $0 < D_\alpha(A) \le D_\alpha\{w : |w(x) - w_0(x)| < \epsilon, \text{ for } x \in \mathcal{X}\}$.

Lemma 3. *Suppose $X \sim Beta(\alpha, \beta)$ with $\alpha, \beta > 0$. Then*

$$E[-\log X] \le \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{2}{\alpha^2} + \log 2.$$

Lemma 4. *For $0 < \beta < 1$, $1/(e\beta) \le \Gamma(\beta) \le 1/\beta + 1/e$.*

Lemma 5. *For $0 < c < a < 1$, $\Gamma(a)/(\Gamma(c)\Gamma(a-c)) \le e^2/2$.*

The proofs of Lemmas 3-5 are relegated to the appendix.

Lemma 6. *Suppose $w \sim D_\alpha$. For t with $\alpha(t) < 1$,*

$$E[\int_{\inf \mathcal{X}}^t f(x|\theta_0)(-\log \frac{w(x)}{w(t)}) d\nu(x)] \le e^2 \int_{\inf \mathcal{X}}^t \frac{f(x|\theta_0)}{\alpha^2(x)} d\nu(x) + F(t|\theta_0) \log 2.$$

Proof. Since $w(x)/w(t) \sim Beta(\alpha(x), \alpha(t) - \alpha(x))$ for $x < t$, Lemmas 3 and 5 yield

$$E[-\log \frac{w(x)}{w(t)}] \le \frac{\Gamma(\alpha(t))}{\Gamma(\alpha(x))\Gamma(\alpha(t) - \alpha(x))} \frac{2}{\alpha^2(x)} + \log 2 \le \frac{e^2}{\alpha^2(x)} + \log 2.$$

Hence,

$$\begin{aligned} E[\int_{\inf \mathcal{X}}^t f(x|\theta_0)(-\log \frac{w(x)}{w(t)}) d\nu(x)] &= \int_{\inf \mathcal{X}}^t f(x|\theta_0) E[-\log \frac{w(x)}{w(t)}] d\nu(x) \\ &\le e^2 \int_{\inf \mathcal{X}}^t \frac{f(x|\theta_0)}{\alpha^2(x)} d\nu(x) + F(t|\theta_0) \log 2. \end{aligned}$$

To simplify the notation, define the following sets. For all $t \in \mathcal{X}$, $\epsilon > 0$ and a vector $v_t = (t_0, \ldots, t_{k+1})$ where $t_0 = t < t_1 < \cdots < t_k < t_{k+1} = \infty$, define

$$\begin{aligned} A_1(t, \epsilon) &= \{w : \sup_{x \ge t} |w_0(x) - w(x)| \le \epsilon\}, \\ A_2(t, \epsilon) &= \{w : \int_{\inf \mathcal{X}}^t f(x|\theta_0) |\log w(x)| d\nu(x) < \epsilon\}, \\ B_1(t, \epsilon, v_t) &= \{w : |\frac{w(t_i) - w(t_{i-1})}{1 - w(t)} - \frac{w_0(t_i) - w_0(t_{i-1})}{1 - w_0(t)}| < \epsilon, i = 1, \ldots, k+1\}, \end{aligned}$$

$$
\begin{aligned}
B_2(t,\epsilon) &= \{w : \int_{\inf \mathcal{X}}^{t} f(x|\theta_0)|\log \frac{w(x)}{w(t)}|d\nu(x) < \epsilon\},\\
B_3(t,\epsilon) &= \{w : |w(t) - w_0(t)| < \epsilon\},\\
K_\epsilon &= \{(\theta, w) : I(f^{w_0}(\cdot|\theta_0), f^{w}(\cdot|\theta)) < \epsilon\}.
\end{aligned}
$$

Lemma 7. *Suppose* $C1$ *holds and* $\mathcal{X} = (a,\infty)$*. Then for all* $\delta > 0$ *and* $t \in \mathcal{X}$ *with* $w_0(t) < 1/2$*, there exist an* $\eta > 0$ *and a vector* $v_t = (t_0,\ldots,t_{k+1})$ *with* $t = t_0 < t_1 < \cdots < t_k < t_{k+1} = \infty$ *such that* $B_1(t,\eta,v_t) \cap B_3(t,\eta) \subset A_1(t,\delta)$*.*

Proof. The proof involves only standard arguments using the continuity of w_0 and monotonicity of w. The details are omitted.

Lemma 8. *Under* $G3$ *and for each* $\epsilon > 0$*, there exists* $\delta > 0$ *s.t.* $\sup_x |w_0(x) - w(x)| < \delta$ *and* $|\theta_0 - \theta| < \delta$ *imply* $|v(\theta_0, w_0) - v(\theta, w)| < \epsilon$*.*

Proof. Suppose $\epsilon > 0$ is given. By G3, we can choose $0 < \delta < \frac{\epsilon}{2}$ such that $|\theta_0 - \theta| < \delta$ implies $\int |f(x|\theta) - f(x|\theta_0)|d\nu(x) < \frac{\epsilon}{2}$. Suppose (θ, w) is such that $|\theta - \theta_0| < \delta$ and $\sup_x |w(x) - w_0(x)| < \delta$. Then,

$$
\begin{aligned}
&|v(\theta,w) - v(\theta_0,w_0)|\\
&\le \int |w(x) - w_0(x)|f(x|\theta)d\nu(x) + \int w_0(x)|f(x|\theta) - f(x|\theta_0)|d\nu(x)\\
&\le \sup_x |w_0(x) - w(x)| + \int |f(x|\theta) - f(x|\theta_0)|d\nu(x) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\end{aligned}
$$

completing the proof.

Lemma 9. *Under* $G3$*, the map* $\theta \to \int_{\mathcal{X}} f^{w_0}(x|\theta_0)|\log \frac{f(x|\theta_0)}{f(x|\theta)}|d\nu(x)$ *is continuous at* θ_0*.*

Proof. Since $w_0(x) \le 1$ for all x, we have

$$
\int f^{w_0}(x|\theta_0)|\log \frac{f(x|\theta_0)}{f(x|\theta)}|d\nu(x) \le \frac{1}{v(\theta_0,w_0)} \int f(x|\theta_0)|\log \frac{f(x|\theta_0)}{f(x|\theta)}|d\nu(x).
$$

The conclusion follows from the assumption that the map $\theta \to \int f(x|\theta_0)|\log \frac{f(x|\theta_0)}{f(x|\theta)}|\, d\nu(x)$ is continuous at θ_0.

Lemma 10. *Suppose* $\mathcal{X} = (a,\infty)$*. Under assumptions* $G1 - G3$ *and* $C1 - C4$*,* $(\pi \times D_\alpha)(K_\epsilon) > 0,$ *for all* $\epsilon > 0$*.*

Proof. We first fix ϵ's and δ's and then sketch the plan for the proof. Suppose $\epsilon > 0$ is given. By Lemma 8 and the continuity of log, there exists $\delta_1 > 0$ such that, for all w and θ satisfying $\sup_x |w_0(x) - w(x)| < \delta_1$ and $|\theta_0 - \theta| < \delta_1$, $|\log v(\theta,w) - \log v(\theta_0,w_0)| < \epsilon$. Now choose $t \in \mathcal{X}$ so that $w_0(t) < \min\{1/2, \delta_1/2\}$,

$\alpha(t) < 1$ and

$$\int_{\inf \mathcal{X}}^{t} f(x|\theta_0)|\log w_0(x)|dx < \epsilon/3 \tag{2}$$

$$\int_{\inf \mathcal{X}}^{t} \frac{f(x|\theta_0)}{\alpha^2(x)}dx < \frac{\epsilon}{12e^2} \tag{3}$$

$$F(t|\theta_0)|\log w_0(t)| < \epsilon/12 \tag{4}$$

$$F(t|\theta_0) < \frac{\epsilon}{12}. \tag{5}$$

Since log is uniformly continuous on every compact interval away from 0 and $w_0(t) > 0$ by C2, we can take $\delta < w_0(t)/2$ (hence $\delta < \delta_1/4$) so that $|\log x - \log y| < \epsilon/3$, for all $x, y \in [w_0(t)/2, 1]$ with $|x - y| < \delta$. By Lemma 9, there exists $\delta_2 < \delta_1$ such that, for θ with $|\theta - \theta_0| < \delta_2$,

$$\int f^{w_0}(x|\theta_0)|\log \frac{f(x|\theta_0)}{f(x|\theta)}|dx < \epsilon.$$

Note that $d\nu(x) = dx$ since ν is Lebesgue measure. Let $A_1 = A_1(t, \delta)$ and $A_2 = A_2(t, \epsilon/3)$, and let $M = 2+1/v(\theta_0, w_0)$. Finally, by Lemma 7, we can choose $0 < \eta < \delta$ and a vector $v_t = (t_0, \ldots, t_{k+1})$ with $t = t_0 < t_1 < \cdots < t_k < t_{k+1}$ so that $B_1(t, \eta, v_t) \cap B_3(t, \eta) \subset A_1$. Let $B_1 = B_1(t, \eta, v_t)$, $B_2 = B_2(t, \epsilon/6)$, and $B_3 = B_3(t, \eta)$.

We now sketch the proof in three steps.

Step 1. $\{(\theta, w) : w \in A_1 \cap A_2, |\theta - \theta_0| < \delta_2\} \subset K_{M\epsilon}$.

After Step 1 is shown, to complete the proof, it suffices to show $D_\alpha(A_1 \cap A_2) > 0$, because the prior $\pi(\theta)$ puts positive mass on every open neighborhood of θ_0. This is not easy to show directly, so we show

Step 2. $B_1 \cap B_2 \cap B_3 \subset A_1 \cap A_2$.

Since B_1, B_2, and B_3 depend only on $w|_{(t,\infty)}$, $w|_{(\inf \mathcal{X}, t]}$, and $w(t)$, respectively, these three sets are independent under D_α by the projection lemma. By the property of Dirichlet distributions, B_1 and B_3 have positive D_α probability. Finally, we show

Step 3. $D_\alpha(B_2) > 0$.

Combining Steps 2-3, we have

$$D_\alpha(A_1 \cap A_2) \geq D_\alpha(B_1 \cap B_2 \cap B_3) = D_\alpha(B_1)D_\alpha(B_2)D_\alpha(B_3) > 0.$$

We complete the proof by showing Steps 1 — 3.

Proof of Step 1. We first show two facts.

a. $\sup_x |w(x) - w_0(x)| < \delta_1$, for all $w \in A_1$.

Because w and w_0 are both nondecreasing positive functions,

$$\sup_{x \leq t} |w_0(x) - w(x)| \leq \max\{w_0(t), w(t)\} \leq w_0(t) + \delta.$$

Finally, because both $w_0(t)$ and δ are smaller than $\delta_1/2$, the result follows.

b. If $w \in A_1 \cap A_2$, then $\int_{\mathcal{X}} f(x|\theta_0)|\log w(x) - \log w_0(x)|dx < \epsilon$.
The integral breaks into three pieces, which can be bounded by

$$\int f(x|\theta_0)|\log w(x) - \log w_0(x)|dx \le \int_{\inf \mathcal{X}}^{t} f(x|\theta_0)|\log w_0(x)|dx$$

$$+ \int_{\inf \mathcal{X}}^{t} f(x|\theta_0)|\log w(x)|dx + \int_{t}^{\infty} f(x|\theta_0)|\log w(x) - \log w_0(x)|dx.$$

By (2), the first term is less than $\epsilon/3$ and the second is less than $\epsilon/3$ because $w \in A_2$. For $w \in A_1$ and $x \ge t$, $|\log w(x) - \log w_0(x)| < \epsilon/3$, because δ was so chosen. Hence, the integral in part **b** is less than ϵ.

Suppose $w \in A_1 \cap A_2$ and $|\theta_0 - \theta| < \delta_2$. Then, since $\sup_x |w_0(x) - w(x)| < \delta_1$ and $|\theta_0 - \theta| < \delta_1$, we have $|\log v(\theta_0, w_0) - \log v(\theta, w)| < \epsilon$. Furthermore,

$$\int f^{w_0}(x|\theta_0)|\log \frac{w_0(x)}{w(x)}|dx \le \frac{1}{v(\theta_0, w_0)} \int f(x|\theta_0)|\log \frac{w_0(x)}{w(x)}|dx < \frac{\epsilon}{v(\theta_0, w_0)}.$$

The first inequality holds because $w_0(x) \le 1$ for all $x \in \mathcal{X}$, and the second inequality holds because of **b**. Combining these inequalities, we have, for $w \in A_1 \cap A_2$ and $|\theta_0 - \theta| < \delta_2$,

$$\begin{aligned} I(f^{w_0}(\cdot|\theta_0), f^{w}(\cdot|\theta)) &= \int f^{w_0}(x|\theta_0) \log \frac{f^{w_0}(\cdot|\theta_0)}{f^{w}(\cdot|\theta)} dx \\ &\le \int f^{w_0}(x|\theta_0)(|\log \frac{w_0(x)}{w(x)}| + |\log \frac{f(x|\theta_0)}{f(x|\theta)}| + |\log \frac{v(\theta, w)}{v(\theta_0, w_0)}|)dx \\ &< \frac{\epsilon}{v(\theta_0, w_0)} + \epsilon + \epsilon = M\epsilon. \end{aligned}$$

This completes the proof of *Step* 1.

Proof of Step 2. Since B_1 and B_3 are chosen so than $B_1 \cap B_3 \subset A_1$, it suffices to show $B_2 \cap B_3 \subset A_2$. We need to show, for $w \in B_2 \cap B_3$, that $\int_{\inf \mathcal{X}}^{t} f(x|\theta_0)|\log w(x)|dx < \epsilon/3$. First, observe that

$$\begin{aligned} \int_{\inf \mathcal{X}}^{t} f(x|\theta_0)|\log w(x)|dx &\le \int_{\inf \mathcal{X}}^{t} f(x|\theta_0)|\log \frac{w(x)}{w(t)}|dx + \int_{\inf \mathcal{X}}^{t} f(x|\theta_0)|\log w(t)|dx \\ &< \epsilon/6 + |\log w(t)|F(t|\theta_0) \\ &\le \epsilon/6 + |\log w_0(t)|F(t|\theta_0) + |\log w_0(t) - \log w(t)|F(t|\theta_0) \\ &< \frac{\epsilon}{6} + \frac{\epsilon}{12} + \frac{\epsilon}{3}\frac{\epsilon}{12} < \frac{\epsilon}{3}. \end{aligned}$$

The second inequality holds because $w \in B_2$; (4) yields the second term in the fourth inequality; (5) together with the definition of δ and the fact that $\eta < \delta$

yields the third term in the fourth inequality. This completes the proof of *Step* 2.

Proof of Step 3.

$$\begin{aligned}D_\alpha(B_2^c) &= D_\alpha\{w : \int_{\inf \mathcal{X}}^{t} f(x|\theta_0)|\log \frac{w(x)}{w(t)}|dx \geq \frac{\epsilon}{6}\} \\ &\leq \frac{6}{\epsilon} E[\int_{\inf \mathcal{X}}^{t} f(x|\theta_0)|\log \frac{w(x)}{w(t)}|dx] \\ &\leq \frac{6}{\epsilon}[e^2 \int_{\inf \mathcal{X}}^{t} \frac{f(x|\theta_0)}{\alpha^2(x)} dx + F(t|\theta_0)\log 2] \\ &< \frac{6}{\epsilon}[\frac{\epsilon}{12} + \frac{\epsilon}{12}] = 1.\end{aligned}$$

The first inequality follows from the Markov inequality, the second from Lemma 6, and the last from (3) and (5). This completes the proof of the lemma.

Lemma 11. *Suppose* $\mathcal{X} = \{a, a+1, \ldots\}$. *Under assumptions* $G1 - G3$ *and* $D1$, $(\pi \times D_\alpha)(K_\epsilon) > 0$ *for all* $\epsilon > 0$.

Proof. Let $\delta > 0$ be given. The Kullback-Leibler divergence can be bounded as follows:

$$\begin{aligned}I(f^{w_0}(\cdot|\theta_0), f^w(\cdot|\theta)) &= \sum_{x=a}^{\infty} f^{w_0}(x|\theta_0) \log \frac{f^{w_0}(x|\theta_0)}{f^w(x|\theta)} \\ &\leq \sum_{x=a}^{\infty} f^{w_0}(x|\theta_0)(|\log \frac{w_0(x)}{w(x)}| + |\log \frac{f(x|\theta_0)}{f(x|\theta)}| + |\log \frac{v(\theta, w)}{v(\theta_0, w_0)}|).\end{aligned}$$

We can choose $\epsilon > 0$ such that, for $(\theta, w) \in A = \{(\theta, w) : \sup_{x \in cX} |w(x) - w_0(x)| < \epsilon,\ |\theta - \theta_0| < \epsilon\}$,

$$|\log \frac{w_0(x)}{w(x)}| < \delta/3 \text{ for all } x \in \mathcal{X}, \tag{6}$$

$$\sum_{x=a}^{\infty} f^{w_0}(x|\theta_0)|\log \frac{f(x|\theta_0)}{f(x|\theta)}| < \delta/3, \tag{7}$$

$$|\log \frac{v(\theta, w)}{v(\theta_0, w_0)}| < \delta/3. \tag{8}$$

The inequality (6) is due to D1 and the fact that log is uniformly continuous on every compact interval away from 0; (7) is due to Lemma 9; and (8) is due to Lemma 8. Hence $I(f^{w_0}(\cdot|\theta_0), f^w(\cdot|\theta)) < \delta$ for $(\theta, w) \in A$. By G2, $D_\alpha(A) > 0$ for all ϵ.

Under the negative binomial sampling scheme, Lemmas 10 and 11 are enough to establish consistency utilizing Schwartz's theorem because, by summing out N

from the likelihood, the posterior is the same as that arising from an i.i.d. sample, i.e., for $A \subset \Theta \times \mathcal{W}$,

$$p(A|O,n) = \frac{\int_A \prod_{i\in O} f^w(x_i|\theta)(\pi \times D_\alpha)(d\theta, dw)}{\int \prod_{i\in O} f^w(x_i|\theta)(\pi \times D_\alpha)(d\theta, dw)}.$$

Under the binomial sampling, the i.i.d. structure does not hold and Schwartz's theorem cannot be applied. However, a closer look at the likelihood suggests that the posterior can be written as one arising from an i.i.d. sample with priors that vary with n, and that Barron's theorem can be applied to establish the consistency of the posterior. Let λ be a prior defined on $\Theta \times \mathcal{W}$, either $(\pi \times D_\alpha)$ or $\int (\pi \times D_{\alpha(\cdot|\eta)})\mu(d\eta)$ where μ is a prior on η, and let N_0 be the true value of N. Define

$$\begin{aligned} m(n|\theta,w) &= \sum_{N=n}^{\infty} \pi_1(N) \binom{N}{n} v(\theta,w)^n (1-v(\theta,w))^{N-n} \\ m(n) &= \int m(n|\theta,w)\lambda(d\theta,dw) \\ p(n) &= \binom{N_0}{n} v(\theta_0,w_0)^n (1-v(\theta_0,w_0))^{N_0-n} \\ \lambda_n(d\theta,dw) &= \frac{m(n|\theta,w)\lambda(d\theta,dw)}{m(n)}. \end{aligned} \tag{9}$$

Then the posterior can be written as, for $A \subset \Theta \times \mathcal{W}$,

$$\begin{aligned} p(A|O,n) &= \frac{\int_A \sum_{N=n}^{\infty} \pi_1(N)\binom{N}{n} v(\theta,w)^n (1-v(\theta,w))^{N-n} \prod_{j=1}^n f^w(x_{i_j}|\theta)\lambda(d\theta,dw)}{\int \sum_{N=n}^{\infty} \pi_1(N)\binom{N}{n} v(\theta,w)^n (1-v(\theta,w))^{N-n} \prod_{j=1}^n f^w(x_{i_j}|\theta)\lambda(d\theta,dw)} \\ &= \frac{\int_A \prod_{j=1}^n f^w(x_{i_j}|\theta)\lambda_n(d\theta,dw)}{\int \prod_{j=1}^n f^w(x_{i_j}|\theta)\lambda_n(d\theta,dw)}. \end{aligned}$$

Using Barron's theorem, it suffices to show that, for all $\epsilon, r > 0$, and for all sufficiently large n, $\lambda_n(K_\epsilon) > e^{-nr}$. To prove this, we need the following lemma.

Lemma 12. *For the prior $\pi_1(N) = N^{-\gamma}$ with $\gamma \geq 1$,*

$$\frac{\nu(\theta,w)^{k+1}}{n(n-1)\ldots(n-k-1)} \leq m(n|\theta,w) \leq \frac{\nu(\theta,w)^{k-1}}{n(n-1)\cdots(n-k+1)},$$

for all sufficiently large n, where k is the integral part of γ.

Proof. We show the second inequality first. Since $k \leq \gamma \leq k+1$,

$$m(n|\theta,w) = \sum_{N=n}^{\infty} \frac{1}{N^\gamma} \binom{N}{n} \nu(\theta,w)^n (1-\nu(\theta,w))^{N-n}$$

$$\begin{aligned}
&\le \sum_{N=n}^{\infty} \frac{1}{N(N-1)\cdots(N-k+1)} \frac{N!}{n!(N-n)!} \nu(\theta,w)^n (1-\nu(\theta,w))^{N-n} \\
&= \frac{\nu(\theta,w)^{k-1}}{n(n-1)\cdots(n-k+1)} \sum_{N=n}^{\infty} \frac{(N-k)!}{(n-k)!(N-n)!} \nu(\theta,w)^{n-k+1} (1-\nu(\theta,w))^{N-n} \\
&= \frac{\nu(\theta,w)^{k-1}}{n(n-1)\cdots(n-k+1)}.
\end{aligned}$$

Since for all sufficiently large n, $N^{k+1} \le N(N-1)\cdots(N-k-1)$, for all $N \ge n$, we have

$$\begin{aligned}
m(n|\theta,w) &\ge \sum_{N=n}^{\infty} \frac{1}{N(N-1)\cdots(N-k-1)} \frac{N!}{n!(N-n)!} \nu(\theta,w)^n (1-\nu(\theta,w))^{N-n} \\
&= \frac{\nu(\theta,w)^{k+1}}{n(n-1)\cdots(n-k-1)} \sum_{N=n}^{\infty} \frac{(N-k-2)!}{(n-k-2)!(N-n)!} \nu(\theta,w)^{n-k-1} (1-\nu(\theta,w))^{N-n} \\
&= \frac{\nu(\theta,w)^{k+1}}{n(n-1)\cdots(n-k-1)}.
\end{aligned}$$

Proof of Theorems 1 and 2. In both theorems, it suffices to show that, for all $\epsilon, r > 0$ and for all sufficiently large n, $\lambda_n(K_\epsilon) \ge e^{-nr}$, where λ_n is given by (9) with $\lambda = \pi \times D_\alpha$. Using Lemma 12,

$$\begin{aligned}
\lambda_n(K_\epsilon) &= \frac{\int_{K_\epsilon} m(n|\theta,w)\lambda(d\theta,dw)}{\int m(n|\theta,w)\lambda(d\theta,dw)} \\
&\ge \frac{\int_{K_\epsilon} \nu(\theta,w)^{k+1}/n(n-1)\cdots(n-k-1)\lambda(d\theta,dw)}{\int \nu(\theta,w)^{k-1}/n(n-1)\cdots(n-k+1)\lambda(d\theta,dw)} \\
&= \frac{1}{(n-k)(n-k-1)} \frac{\int_{K_\epsilon} \nu(\theta,w)^{k+1}\lambda(d\theta,dw)}{\int \nu(\theta,w)^{k-1}\lambda(d\theta,dw)},
\end{aligned}$$

where k is the integral part of γ. Because the first factor $1/[(n-k)(n-k-1)]$ decreases at a polynomial rate and the second factor is a positive quantity (due to Lemmas 10 and 11) and independent of n, we are done.

Proof of Theorem 3. The proof is the same as that of Theorems 1 and 2, except that $\lambda = \int (\pi \times D_{\alpha(\cdot|\eta)})\mu(d\eta)$.

Acknowledgements

This work was supported, in part, by the National Science Foundation under Grants DMS-9303556 and DMS-9802261, and formed part of the author's Ph.D. thesis at Purdue University. The author thanks the Associate Editor, the referees, and especially the Editor for careful reading and constructive suggestions.

Appendix

Proof of Lemma 3. The expectation, without the normalizing constant, breaks into

$$\int_0^1 (-\log x)x^{\alpha-1}(1-x)^{\beta-1}dx = \int_0^{1/2}(-\log x)x^{\alpha-1}(1-x)^{\beta-1}dx + \int_{1/2}^1(-\log x)x^{\alpha-1}(1-x)^{\beta-1}dx.$$

The first integral can be bounded using the fact that $(1-x)^{\beta-1} \le 2$ for $x < 1/2$ and $\beta > 0$;

$$\int_0^{1/2}(-\log x)x^{\alpha-1}(1-x)^{\beta-1}dx \le 2\int_0^{1/2}(-\log x)x^{\alpha-1}dx \le \frac{2}{\alpha^2}.$$

The second integral can be bounded similarly, using the inequality $-\log x \le \log 2$ for $1/2 \le x \le 1$:

$$\int_{1/2}^1(-\log x)x^{\alpha-1}(1-x)^{\beta-1}dx \le \log 2\int_{1/2}^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \log 2\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The above two inequalities yield the result.

Proof of Lemma 4. Write

$$\Gamma(\beta) = \int_0^\infty x^{\beta-1}e^{-x}dx = \int_0^1 x^{\beta-1}e^{-x}dx + \int_1^\infty x^{\beta-1}e^{-x}dx.$$

Note that

$$\frac{1}{e\beta} = \int_0^1 x^{\beta-1}e^{-1}dx \le \int_0^1 x^{\beta-1}e^{-x}dx \le \int_0^1 x^{\beta-1}e^{-0}dx = \frac{1}{\beta}$$

and, for $0 < \beta < 1$,

$$\int_1^\infty x^{\beta-1}e^{-x}dx \le \int_1^\infty 1^{\beta-1}e^{-x}dx = \frac{1}{e}.$$

Combining the above inequalities, we get the result.

Proof of Lemma 5. Using Lemma 4 and the fact that $1/a > 1/e$, we have

$$\frac{\Gamma(a)}{\Gamma(c)\Gamma(a-c)} \le \frac{\frac{1}{a}+\frac{1}{e}}{\frac{1}{ec}\frac{1}{e(a-c)}} \le e^2c(a-c)\frac{2}{a} \le 2e^2\frac{c}{a}(1-\frac{c}{a}).$$

Finally, because $p(1-p) \le 1/4$ for all $0 \le p \le 1$, we get the conclusion.

References

Barron, A. R. (1986). Comments on "On the consistency of Bayes estimates". *Ann. Statist.* **14**, 26-30.

Barron, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report, Department of Statistics, University of Illinois, Champaign, IL.

Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27**, 536-561.

Bayarri, M. J. and DeGroot, M. H. (1990). Selection models and selection mechanisms. In *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (Edited by S. Geisser, J. S. Hodges, S. J. Press and A. E. Zellner), 211-227. North-Holland/Elsevier, Amsterdam.

Bayarri, M. J. and DeGroot, M. H. (1992). A BAD view of weighted distributions and selection models. In *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting* (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 17-29. Clarendon Press, Oxford.

Cook, R. D. and Martin, F. B. (1974). A model for quadrat sampling with visibility bias. *J. Amer. Statist. Assoc.* **69**, 345-349.

Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21**, 903-923.

Diaconis, P. and Freedman, D. (1986). On inconsistent Bayes estimates of location. *Ann. Statist.* **14**, 68-87.

Diaconis, P. and Freedman, D. (1993). Nonparametric binary regression: a Bayesian approach. *Ann. Statist.* **21**, 2108-2137.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Consistency issues in Bayesian nonparametrics. In *Asymptotic Nonparametrics and Time Series: A Tribute to Madan Lal Puri* (Edited by Subir Ghosh), 639-667. Marcel Dekker, New York.

Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statist. Sci.* **3**, 109-117.

Lee, J. and Berger, J. O. (1999). Semiparametric Bayesian analysis of selection models. Discussion Paper, ISDS, Duke University.

Maddala, G. S. (1977). Self-selectivity problems in econometric models. In *Applications of Statistics* (Edited by P. R. E. Krishnaiah), 351-366. North-Holland/Elsevier, Amsterdam.

Meisner, J. and Demirmen, F. (1981). The creaming method: a Bayesian procedure to forecast future oil and gas discoveries in mature exploration provinces. *J. Roy. Statist. Soc. Ser. A* **144**, 1-31.

Nair, V. N. and Wang, P. C. C. (1989). Maximum likelihood estimation under a successive sampling discovery model. *Technometrics* **31**, 423-436.

Patil, G. P. and Rao, C. R. (1977). The weighted distributions: a survey of their applications. In *Applications of Statistics* (Edited by P. R. E. Krishnaiah), 383-406. North-Holland/Elsevier, Amsterdam.

Schwartz, L. (1965). On Bayes procedures. *Zeit. Wahrsch. Verw. Gebiete* **4**, 10-26.

Silliman, N. P. (1997a). Non-parametric classes of weight functions to model publication bias. *Biometrika* **84**, 909-918.

Silliman, N. P. (1997b). Hierarchical selection models with applications in meta-analysis. *J. Amer. Statist. Assoc.* **92**, 926-936.

Sethuraman, J. (1983). Unpublished Lecture Note on Dirichlet Processes. CBMS Conference Board on Mathematics.

Sun, J. and Woodroofe, M. B. (1997). Semi-parametric estimates under biased sampling. *Statist. Sinica* **7**, 545-576.

West, M. (1994). Discovery sampling and selection models. In *Statistical Decision Theory and Related Topics V* (Edited by S. S. Gupta and J. O. Berger), 221-235. Springer-Verlag, New York.

West, M. (1996). Inference in successive sampling discovery models. *J. Econometrics* **75**, 217-238.

Pennsylvania State University, 326 Thomas Building, University Park, PA 16802, U.S.A.

E-mail: leej@stat.psu.edu

(Received July 1999; accepted January 2001)